MINING AUXILIARY OBJECTS FOR TRACKING BY MULTIBODY GROUPING

Ming Yang, Ying Wu

Dept. of EECS, Northwestern Univ.
Evanston, IL 60208
{mya671,yingwu}@ece.northwestern.edu

Shihong Lao

Sensing & Control Tech. Lab, OMRON
Kyoto 619-0283, Japan
lao@ari.ncl.omron.co.jp

ABSTRACT

On-line discovery of some auxiliary objects to verify the tracking results is a novel approach to achieving robust tracking by balancing the need for strong verification and computational efficiency. However, the applicability and effectiveness of this approach highly depend on how to reliably validate the motion correlation between the target and the auxiliary objects so as to estimate the motion model. In this paper, we extend the algorithm of mining auxiliary objects for tracking by incorporating multibody grouping to detect the motion correlation and estimate the motion model, which imposes more general motion correlation constraints. The proposed method discovers the auxiliary objects that exhibit strong affine motion correlation and estimates the closed-form affine models. The proposed tracking algorithm shows good performance in real-world test sequences.

Index Terms— Visual tracking, auxiliary objects, multibody grouping, belief propagation.

1. INTRODUCTION

Visual tracking gains much research interests due to its diverse applications in video analysis. One fundamental obstacle against long-duration robust tracking is the lack of efficient verification means. Thus, the tracker may either drift gradually and unconsciously, or a short term invalidation of the target observation model for just several frames, e.g. the target moves out of the image boundary shortly or severe occlusion happens, may cause tracking failure.

The challenge originates from the contradictive requirements on the target observation model. It has to be powerful and robust for tough situations, e.g. clutter backgrounds, illumination and view changes, and partial occlusions, while at the same time to be efficient for real-time processing. Since in real-world applications, the appearances of the target and/or the environment may be non-stationary, it is generally very difficult, if not impossible, to obtain simple visual invariants of the target for tracking. Thus, there have been two primary means to deal with this dilemma, 1) on-line adaptation, e.g. switching among multiple observation models [5], selecting discriminative features [3], incrementally updating observation models [6, 1], and 2) learning a comprehensive target model by off-line training [8] which is tantamount to the ultimate verification, i.e. object recognition. However, on-line adaptation is risky if no other supervised mechanisms to prevent model drifting, and training off-line tends to be very computational demanding and unable to cover all possible variations of the target appearances.

A new approach called intelligent collaborative tracking (ICT) [10] has been proposed to handle this dilemma by taking advantage of the so called *auxiliary objects* which have temporary motion correlation with the targets and are discovered on-the-fly to help verifying the tracking results in a collaborative way. The intuition is that the target is seldom isolated and it is likely that there exist some informative image regions, i.e. auxiliary objects, that have short-term motion correlation with it. Such auxiliary objects cannot be specified in advance or be trained off-line because their appearances and motion change in video. Some sample auxiliary objects are shown in Fig 1, as the yellow boxes indicate the target (i.e. the head) and the dash red boxes show some sample auxiliary objects discovered on-the-fly.

Fig. 1. Some sample auxiliary objects of the target head. They need to be discovered on-the-fly in the tracking process.

The critical point in the ICT is to reliably determine whether one candidate auxiliary object has a strong short-term motion correlation with the target or not. In [10], thresholds on the variances of relative distances and relative scales are used as criteria which are ad hoc and may overlook some general motion correlations. In this paper we address this important issue by employing multibody grouping [9] to discover the poten-

tial multibody structure from motion and estimate the affine motion model through noise subspace analysis. By performing eigenvalue decomposition on the trajectories of the candidate auxiliary objects and the target in a time window, we can check if there is a stable affine motion model. If so, by identifying the noise subspace, the affine motion model can be estimated in a closed-form.

2. MINING AUXILIARY OBJECTS

2.1. Auxiliary objects

Auxiliary objects (AOs) are those that can help tracking due to their strong short-term motion correlation with the target. In fact, it is not necessary for an AO to be a semantic object. In the tracking scenario, it refers to an informative image region or image feature. Specifically, an auxiliary object should satisfy three properties at least in a short time interval: (1) persist co-occurrence with the target, (2) consistent motion correlation with the target, and (3) easy to track.

To discover such AOs during the tracking process, first we need to identify some candidate AOs. Since image regions, if selected properly, can be reliably and efficiently tracked, for example, by the Mean-shift algorithm [4], we employ color regions as candidate auxiliary objects to satisfy property (1). For each frame, we perform efficient quad-tree color segmentation to obtain some color regions and establish their correspondences in consecutive frames by matching their color histograms. By thresholding their frequencies in a time window, a subset of such color regions are selected as candidate auxiliary objects which satisfy property (2) (i.e. persist co-occurrent with the target). Then, the key question is to tell whether they bear strong and consistent motion correlation with the target or not, and how to integrate them in the tracking process.

2.2. Mining by multibody grouping

The motion correlation between two moving objects can be very complicated and non-linear, but generally linear motion models are more feasible to process. In this paper, we extend the simple translational model in [10] to a more general affine motion model. When the points on two objects have affine motion relation, they must reside in a linear subspace [9]. Thus, identifying this subspace will lead to the estimation of the affine motion model.

At time t, one candidate auxiliary object O is represented as $\mathbf{x}_t = \{u_t^x, v_t^x\}^T$ and $\{s_t^u, s_t^v\}$ where (u_t^x, v_t^x) are the coordinates of the center of O and s_t^u and s_t^v are the scales, respectively. Similarly the target T can be represented as $\mathbf{y}_t = \{u_t^y, v_t^y\}^T$ and $\{s_t^u, s_t^v\}$. If O and T co-occur in a short period and have stable motion correlation, then O can be claimed as an auxiliary object. So the goal is to evaluate whether O and T have strong motion correlation in time window $[t-N, t]$ given the trajectories of $\mathbf{y}_t$ and $\mathbf{x}_t$ within this time window.

Assume an affine motion model between auxiliary object O and the target T, which is specified by a 2×2 matrix $\mathbf{A}$ and a translation vector $\mathbf{b} = \{u_b, v_b\}^T$, as

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \mathbf{b}. \tag{1}$$

Subtract the mean $\bar{\mathbf{y}}$ of $\mathbf{y}$ and $\bar{\mathbf{x}}$ of $\mathbf{x}$ in the time window $[t-N, t]$ and take the noise into consideration, the relation between O and T can be expressed with $\tilde{\mathbf{y}}_t = \mathbf{y} - \bar{\mathbf{y}}$ and $\tilde{\mathbf{x}}_t = \mathbf{x} - \bar{\mathbf{x}}$, as

$$\tilde{\mathbf{y}}_t = \mathbf{A}\tilde{\mathbf{x}}_t + \mathbf{n}, \tag{2}$$

where $\mathbf{n}$ is a zero mean white noise with $E[\mathbf{n}\mathbf{n}^T] = \sigma^2\mathbf{I}$.

If we stack $\tilde{\mathbf{y}}_t$ and $\tilde{\mathbf{x}}_t$, the covariance matrix $\mathbf{C}$ can be expressed as

$$\mathbf{C} = E[\begin{pmatrix} \tilde{\mathbf{y}}_t \\ \tilde{\mathbf{x}}_t \end{pmatrix} (\tilde{\mathbf{y}}_t^T, \tilde{\mathbf{x}}_t^T)]. \tag{3}$$

It is clear that $rank(\mathbf{C}) \leq 2$ if there is no noise (i.e. $\mathbf{n} = 0$). This rank deficiency property is important in detecting the subspace due to motion correlation. In reality, because $\mathbf{n} \neq 0$, $\mathbf{C}$ is likely to have a full rank. Since the noise is additive, it is easy to prove that the 4D space spanned by $\left(\tilde{\mathbf{y}}_t^T, \tilde{\mathbf{x}}_t^T\right)$ is a direct sum of a signal subspace and a noise subspace. The signal subspace is up to rank 2 and corresponds to the large eigenvalues of $\mathbf{C}$, and the noise subspace corresponds to the smallest eigenvalues (i.e. σ). Therefore, we can check and threshold the eigenvalues to identify those subspaces.

Denote the estimated covariance matrix by $\hat{\mathbf{C}}$ and the covariance matrix of $\tilde{\mathbf{x}}$ by $\hat{\mathbf{C}}^x$, and we have

$$\hat{\mathbf{C}} = \sum_{i=0}^{N} \begin{pmatrix} \tilde{\mathbf{y}}_{t-i} \\ \tilde{\mathbf{x}}_{t-i} \end{pmatrix} (\tilde{\mathbf{y}}_{t-i}^T, \tilde{\mathbf{x}}_{t-i}^T) = \begin{pmatrix} \mathbf{A}\hat{\mathbf{C}}^x\mathbf{A}^T + \sigma^2 & \mathbf{A}\hat{\mathbf{C}}^x \\ \hat{\mathbf{C}}^x\mathbf{A}^T & \hat{\mathbf{C}}^x \end{pmatrix}. \tag{4}$$

Performing eigenvalue decomposition on $\hat{\mathbf{C}}$,

$$\hat{\mathbf{C}} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}, \tag{5}$$

we obtain the sorted eigenvalues $\{\lambda_1, \cdots, \lambda_4\}$. If there are more than 2 eigenvalues $\lambda_j^2 \gg \sigma^2$, this candidate is not an auxiliary object since its motion and the target's are not in one subspace.

$$\#\text{ of } \{\lambda_j^2 \gg \sigma^2\} \begin{cases} > 2 & \text{NOT AO} \\ <= 2 & \text{AO} \end{cases}. \tag{6}$$

If the candidate is an auxiliary object, we can estimate its affine matrix $\mathbf{A}$ with the property that the noise subspace is orthogonal to the signal subspace. The least two eigenvectors correspond to the noise subspace of $\hat{\mathbf{C}}$ are denoted as

$$\begin{pmatrix} q_{31} & q_{41} \\ q_{32} & q_{42} \\ q_{33} & q_{43} \\ q_{34} & q_{44} \end{pmatrix}.$$

Substitute back to $\hat{\mathbf{C}}$, the 2×2 matrix $\mathbf{A}$ can be solved by

$$\mathbf{A}^T \begin{pmatrix} q_{31} & q_{41} \\ q_{32} & q_{42} \end{pmatrix} + \begin{pmatrix} q_{33} & q_{43} \\ q_{34} & q_{44} \end{pmatrix} = 0. \quad (7)$$

Then, the translation vector $\mathbf{b}$ is obtained with $\bar{\mathbf{y}}$, $\bar{\mathbf{x}}$, and $\mathbf{A}$. This method gives an effective detection of auxiliary objects and efficient estimation of their affine motion models.

3. COLLABORATIVE TRACKING

The above mining process automatically discovers a set of auxiliary objects. How to fuse the motion information of these auxiliary objects to help tracking is also a critical problem. With the mining results, a random field can be learned to model the relation among the target and the auxiliary objects. Not causing confusion, we omit the subscript of time t for short, we denote the motion of the target by $\mathbf{y}$ and those of the auxiliary object by $\mathbf{x}_k, k = 1, \ldots, K$, where K is the number of auxiliary objects. They constitute a random field. Each pair of the target and an auxiliary object $\mathbf{x}_k$ bears a pairwise potential $\psi_{k0}(\mathbf{x}_k, \mathbf{y})$,

$$\psi_{k0}(\mathbf{x}_k, \mathbf{y}) \propto e^{-\frac{(\mathbf{y}-\mathbf{A}_k\mathbf{x}_k-\mathbf{b}_k)^T(\mathbf{y}-\mathbf{A}_k\mathbf{x}_k-\mathbf{b}_k)}{2\sigma^2}}, \quad (8)$$

where σ^2 is derived from the small eigenvalues of $\mathbf{C}$ in Eq.2. In many cases, auxiliary objects share almost the same motion as the target, e.g., the torso and the target head. Therefore, we can use a Gaussian distribution to characterize those potentials. The mean of the Gaussian is given by $\mathbf{A}_k$ and $\mathbf{b}_k$, which is the affine motion model estimated for the kth AO.

Certainly, in the tracking scenario, such a random field is hidden and need to be inferred from image evidence. We formulate this problem under a Markov network with a special topology, as shown in Fig. 2, where we only assume pair-wise connections between the target $\mathbf{y}$ and the auxiliary object $\mathbf{x}_k$ and there are no connections among auxiliary objects. Each of them is associated with its image evidence $\mathbf{z}_k$. We denote $\mathbf{Z} = \{\mathbf{z}_k, k = 0, \ldots, K\}$, where $\mathbf{z}_0$ is the observation of $\mathbf{y}$. The core of tracking is to estimate the posteriors $p(\mathbf{y}|\mathbf{Z})$ of the target and $p(\mathbf{x}_k|\mathbf{Z}), k = 1, \ldots, K$, for the auxiliary objects.

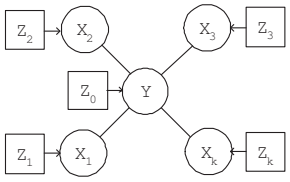

Fig. 2. Star topology random field.

For such a singly connected network, a belief propagation algorithm [7] with 2-step message passing gives the exact estimates of the posteriors.

$$p(\mathbf{y}|\mathbf{Z}) \propto \hat{p}_0(\mathbf{y}|\mathbf{Z}) \prod_k m_{k0}(\mathbf{y}), \quad (9)$$

$$m_{k0}(\mathbf{y}) = \int_{x_k} \hat{p}_k(\mathbf{x}_k|\mathbf{Z}) \psi_{k0}(\mathbf{x}_k, \mathbf{y}) d\mathbf{x}_k, \quad (10)$$

$$p(\mathbf{x}_k|\mathbf{Z}) \propto \hat{p}_k(\mathbf{x}_k|\mathbf{Z}) m_{0k}(\mathbf{x}_k) \quad k = 1, \ldots, K, \quad (11)$$

$$m_{0k}(\mathbf{x}_k) = \int_y \hat{p}_0(\mathbf{y}|\mathbf{Z}) \prod_{\mathbf{x}_i \setminus \mathbf{x}_k} m_{i0}(\mathbf{y}) d\mathbf{y}, \quad (12)$$

where $m_{k0}(\mathbf{y})$ represents the message passed from the kth auxiliary object to the target and $m_{0k}(\mathbf{x}_k)$ is the message from the target to the kth auxiliary object.

If the collaborative tracking result of target is not consistent with its evidence and the auxiliary objects, we assert that the target is experiencing occlusion or drift, and stop the mining process temporarily. If one auxiliary object is not consistent with all the others, we simply exclude this auxiliary object from fusion. Please refer to [10] for the details.

4. EXPERIMENTAL RESULTS

We evaluate the improved ICT algorithm in a head tracking system, where the head tracker is a contour-based elliptical tracker similar to [2], and the auxiliary trackers are Mean-shift trackers. In our experiments, we compare the proposed ICT algorithm with the single contour tracker and the popular Mean-shift tracker [4].

The motion parameters $\{u, v, s_u, s_v\}$ to be recovered include the location (u, v) and the scales s_u and s_v. The quad-tree color segmentation and the mean-shift tracker work in the normalized R-G color space with 32×32 bins. Without code optimization, our C++ implementation of ICT comfortably runs around 10 fps on average on Pentium 3G for 320×240 images depending on the number of the auxiliary objects.

For a quantitative evaluation, we manually labeled the ground truth of the sequences `birthday kid` for 1460 frames. The evaluation criteria of tracking error are based on the relative position errors between the tracking result and that of the ground truth, and the relative scale normalized by the ground truth scale. Ideally, the position differences should be around 0, and the relative scales 1. Note, since the Mean-shift tracker loses tracking after about 500 frames with too large error, we only show its results for 500 frames.

Some key frames are shown in Fig. 4 [1]. The first and second rows show the results of the single Mean-shift tracker and the single contour tracker respectively, where the solid-yellow box indicates the location of the head. The tracking results of the improved ICT are shown in the 3rd row as highlighted solid-yellow box, and the dash-red boxes are the auxiliary object trackers.

[1] All the faces in this paper were mosaicked for privacy protection.

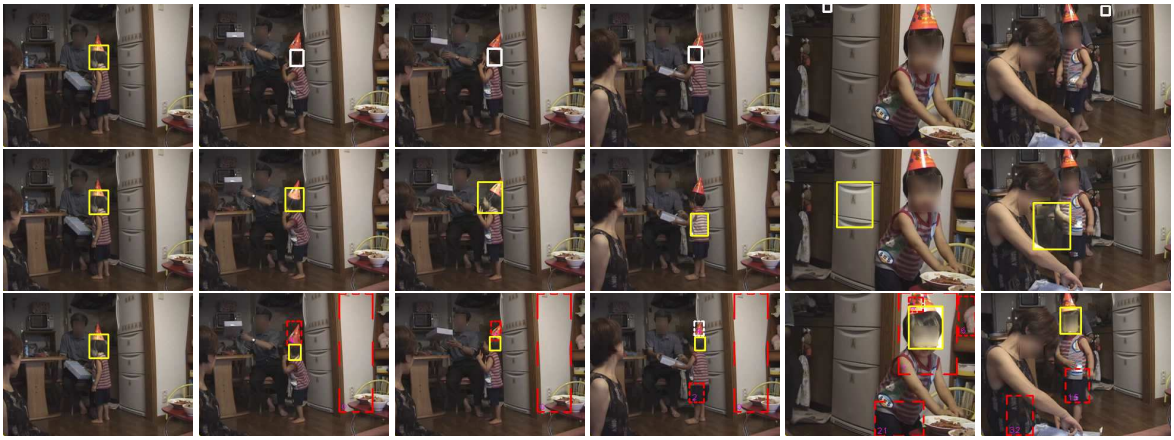

Fig. 4. Frame # 0, 72, 93, 170, 578 and 1455 of `birthday kid,1460 frames`. (1st row) the single Mean-shift tracker, (2nd row) the single contour tracker, (3rd row) the improved ICT tracker.

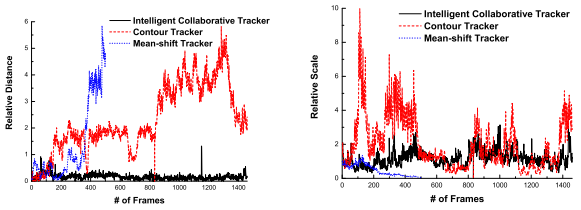

Fig. 3. Quantitative comparison: (left) position errors, (right) scale errors, [`birthday kid,1460 frames`].

Since the target head experiences large out-of-plane rotation and the appearances change greatly, Mean-shift tracker drifts after the kid turns around after frame 400. For the contour tracker, when the rear head is in the dark background, no good observation is available around the head so the contour tracker drifts to the torso and other elliptical regions, and is unable to recover. For the improved ICT tracker, with the help of the auxiliary objects, the tracker either keeps tracking in the tough situations or recovers from drifting in several frames. Note the auxiliary objects discovered can be some objects with inherent relations with the target, such as the hat and short pant, or just happening to have temporary relations, such as the refrigerator or the gift box. This real-world sequence demonstrates the advantage of the auxiliary objects for long-duration tracking.

5. CONCLUSION

In this paper, we improve the intelligent collaborative tracking by incorporating multibody grouping to detect the motion correlation between the target and the auxiliary objects. By incorporating more general affine motion models, the auxiliary objects can be identified reliably and contribute more to long-duration tracking in practical applications.

Acknowledgments

This work was supported in part by National Science Foundation Grants IIS-0347877 and IIS-0308222.

6. REFERENCES

[1] Shai Avidan. Ensemble tracking. In *CVPR*, volume 2, pages 494 – 501, June 20-26, 2005.

[2] Stan Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *CVPR*, pages 232 – 237, Santa Barbara, CA, June 23-25, 1998.

[3] Robert T. Collins and Yanxi Liu. On-line selection of discriminative tracking features. In *ICCV*, volume 2, pages 346–352, Nice, France, October 13-16, 2003.

[4] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. Real-time tracking of non-rigid objects using mean shift. In *CVPR*, volume 2, pages 142–149, Hilton Head Island, South Carolina, June 13-15, 2000.

[5] Greg Hager and Peter Belhumeur. Real-time tracking of image regions with changes in geoemtry and illumination. In *CVPR*, pages 403–410, San Francisco, June 18-20, 1996.

[6] Jongwoo Lim, David Ross, Ruei-Sung Lin, and Ming-Hsuan Yang. Incremental learning for visual tracking. In *NIPS*, pages 801–808, Vancouver, Canada, December 13-18, 2004.

[7] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc., 1998.

[8] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, volume I, pages 511 – 518, Hawaii, Dec. 2001.

[9] Ying Wu, Zhengyou Zhang, Thomas S. Huang, and John Y.Lin. Multibody grouping via orthogonal subspace decomposition. In *CVPR*, volume 2, pages 252 – 257, Hawaii, December 11-13, 2001.

[10] Ming Yang, Ying Wu, and Shihong Lao. Intelligent collaborative tracking by mining auxiliary objects. In *CVPR*, volume 1, pages 697 – 704, NYC, June 17-22, 2006.